Automated Event Service

Tom Clune
CISTO
NASA GSFC

Our Extended Team

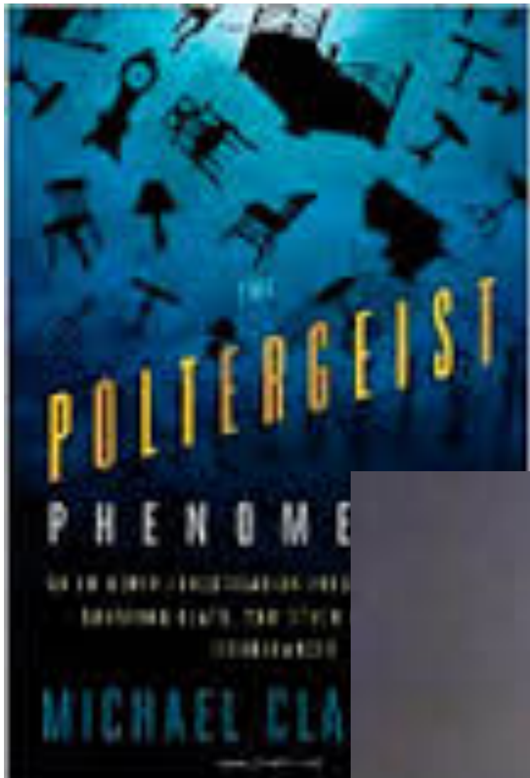
AES: K.-S. Kuo[1,2], J.A. Rushing[3], R. Ramachandran[4], A. Lin[3], G. Fekete[5,1], K. Doan[9,1], R. Tucker[3]

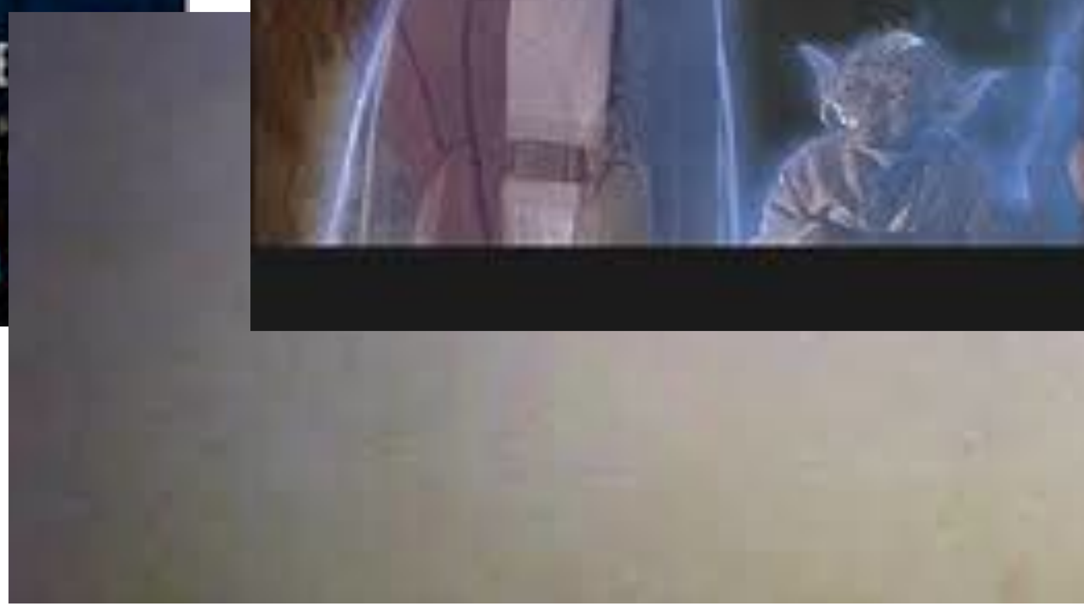
PROBE: K.-S. Kuo[1,2], M. Bauer[8,1], G. Schmidt[1], A. Oloso[6,1], G. Fekete[5,1]

AWS/MODB: K.-S. Kuo[1,2], M. Schneider[10], R. Linan[7,1], A. Oloso[6,1]

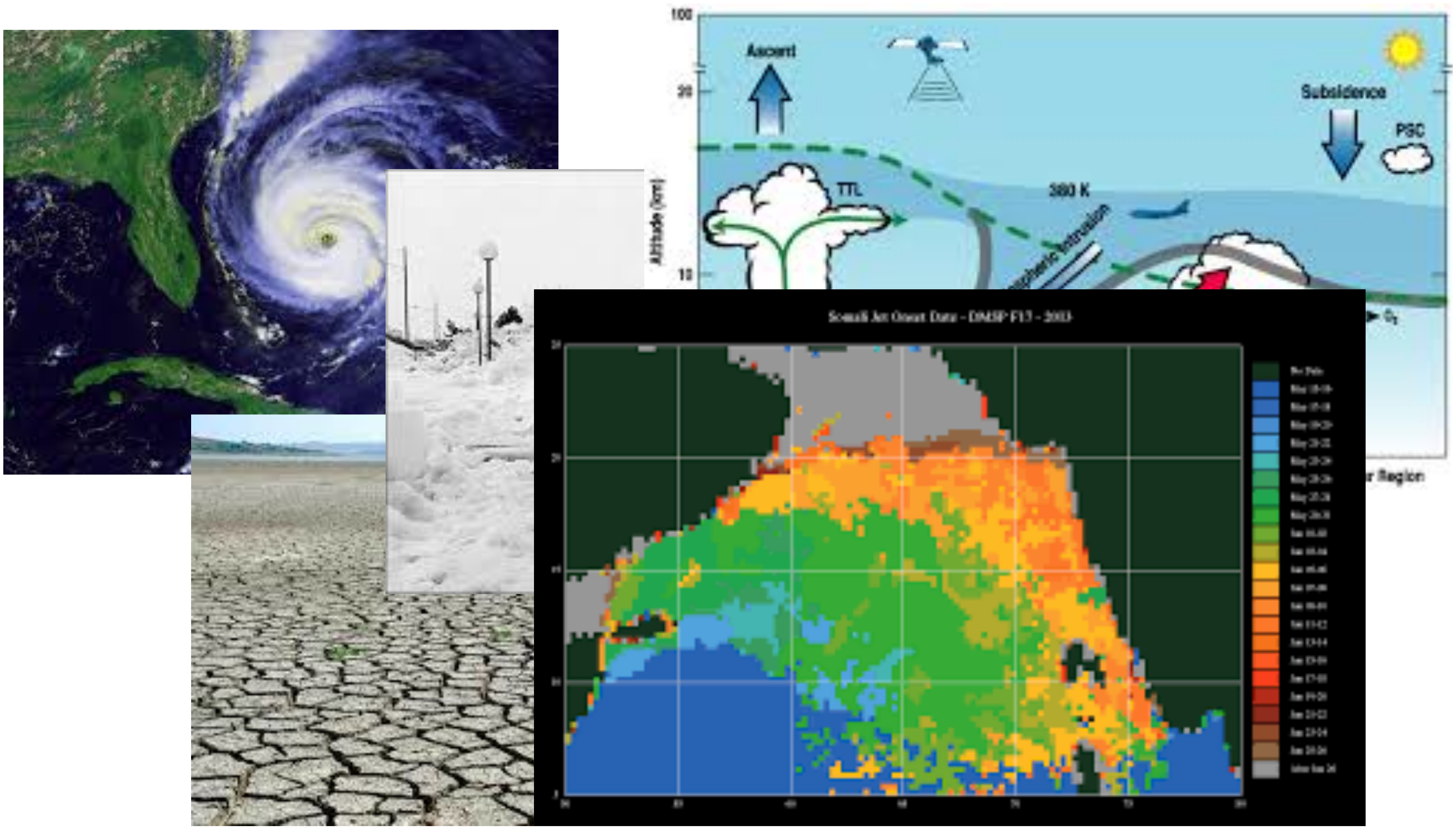
1. NASA GSFC
2. Bayesics, LLC
3. University of Alabama-Huntsville
4. NASA MSFC
5. Computer Science Corporation
6. Science Systems and Applications, Inc.
7. Navteca Inc.
8. Columbia University
9. University of Maryland
10. University of Florida

What if you want to find all instances of some phenomenon?





What if you want to find all instances of some (**<u>Earth Science</u>**) phenomenon?



Searching the ~~Meta~~data

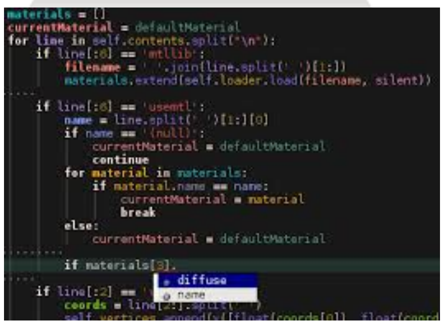
Data center catalogs generally support searches on metadata (where, when, which instrument) rather than on data (temperature, wind, …).

The Usual Process

(1) Download the relevant data collection

(1b) And wait …

(1c) Procure more disk; upgrade network; wait



(2) Write script that searches single time slice

(2b) Don’t share script with others

(3) Loop over all time slices – accumulate

(3b) And wait …

Automated Event Service

- Enable **systematic** identification of **investigator-defined** Earth science events from reanalysis and satellite data.
  - Addresses significant portion of ES research;
  - Reduces duplication of effort among research teams;
  - Improves ROI for NASA data and compute resources.
- Promote affinity between computing and data resources
  - Move the computation **<u>not</u>** the data.
- Improve interactive data exploration and analysis.

24th century technology

AES Major Features

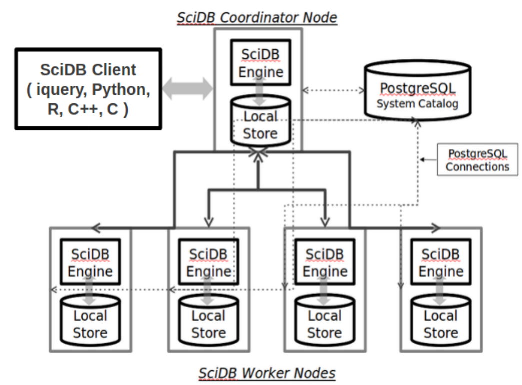
- Custom user-defined operators (UDOs)
  - E.g., Connected Component Labeling (CCL)
- Event Specification Language (ESL)
  - Allowing scientists to express their using Python rather than low-level SQL.
- Collaboration via the Collaborative Workbench (CWB).
  - Event definitions and search results can be shared and modified.
- Scalable parallel performance.
- Web service
  - Allows AES to be embedded within other applications.

~~Data Mining?~~ ~~Data Analytics?~~

Big Data Technology: SciDB

An all-in-one ***data management*** and advanced ***analytics platform***:

- Complex analytics inside a next-generation parallel ***array*** database,
  - *i.e.* **not** *row-based* or *column-based* like RDBMS's based on *table* data model
  - Array Functional Language (AFL)
  - Array Query Language (AQL)
- Based on the "shared nothing architecture" for data parallelism,
- Data versioning and provenance to support science applications, and
- Open source (currently in beta).



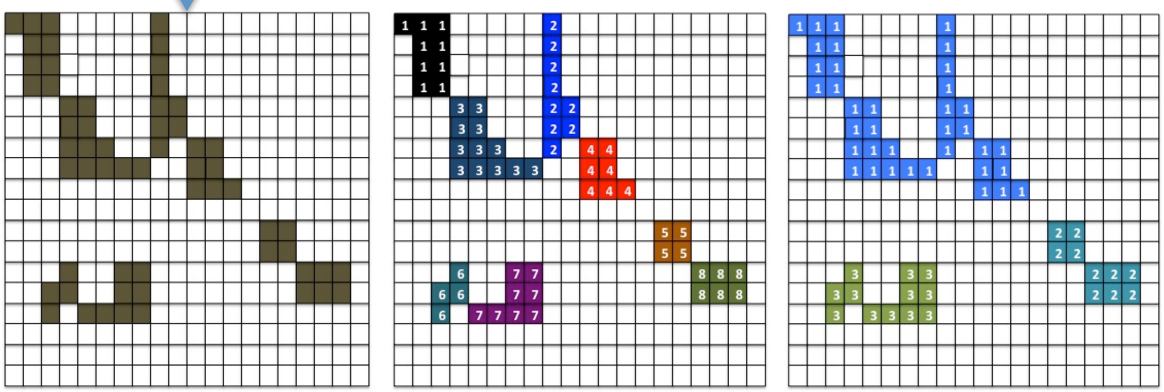
Faster than Hadoop (MapReduce)

2-10x in most benchmarks that we have performed.

Connected Component Labeling

Point-wise query computes a mask for locations which satisfy phenomenon criteria

A. Oloso



a. Filtered data (Mask)

b. CCLs for 4-connectivity

c. CCLs for 8-connectivity

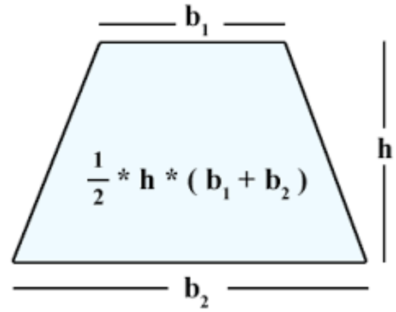
Pixels that are connected in space and/or time are associated with same event.

AES supports CCL in up to 4 dimensions.

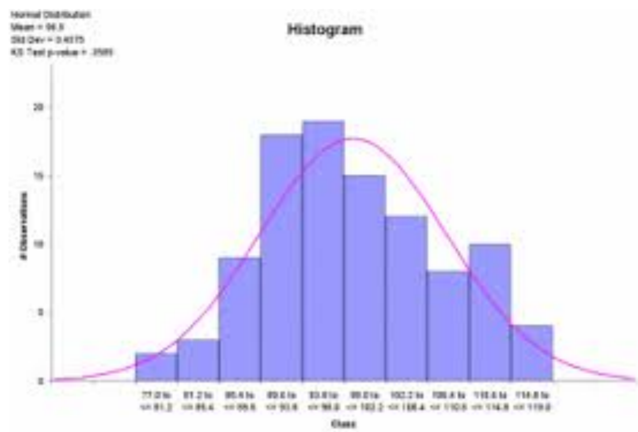
Additional filters can prune events below a certain size or duration.

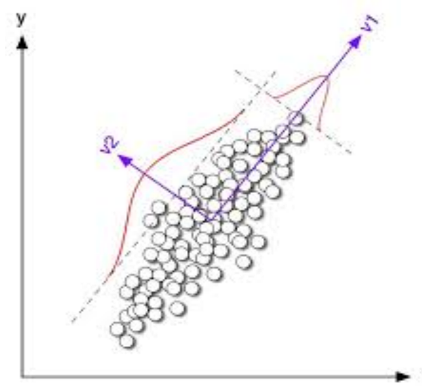
Characterizing Events

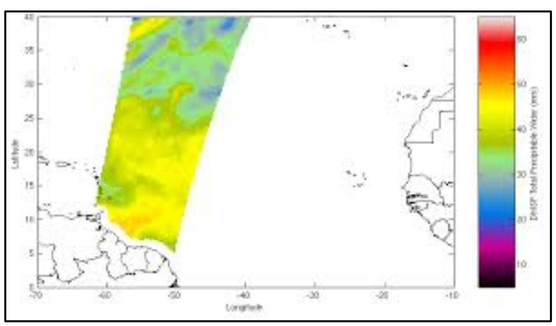
AES annotates events with user-specified per-event properties such as: Area, Volume, Duration, centroid, min temperature, max wind, etc.



Event statistics such as histograms and PCA can be generated from such properties

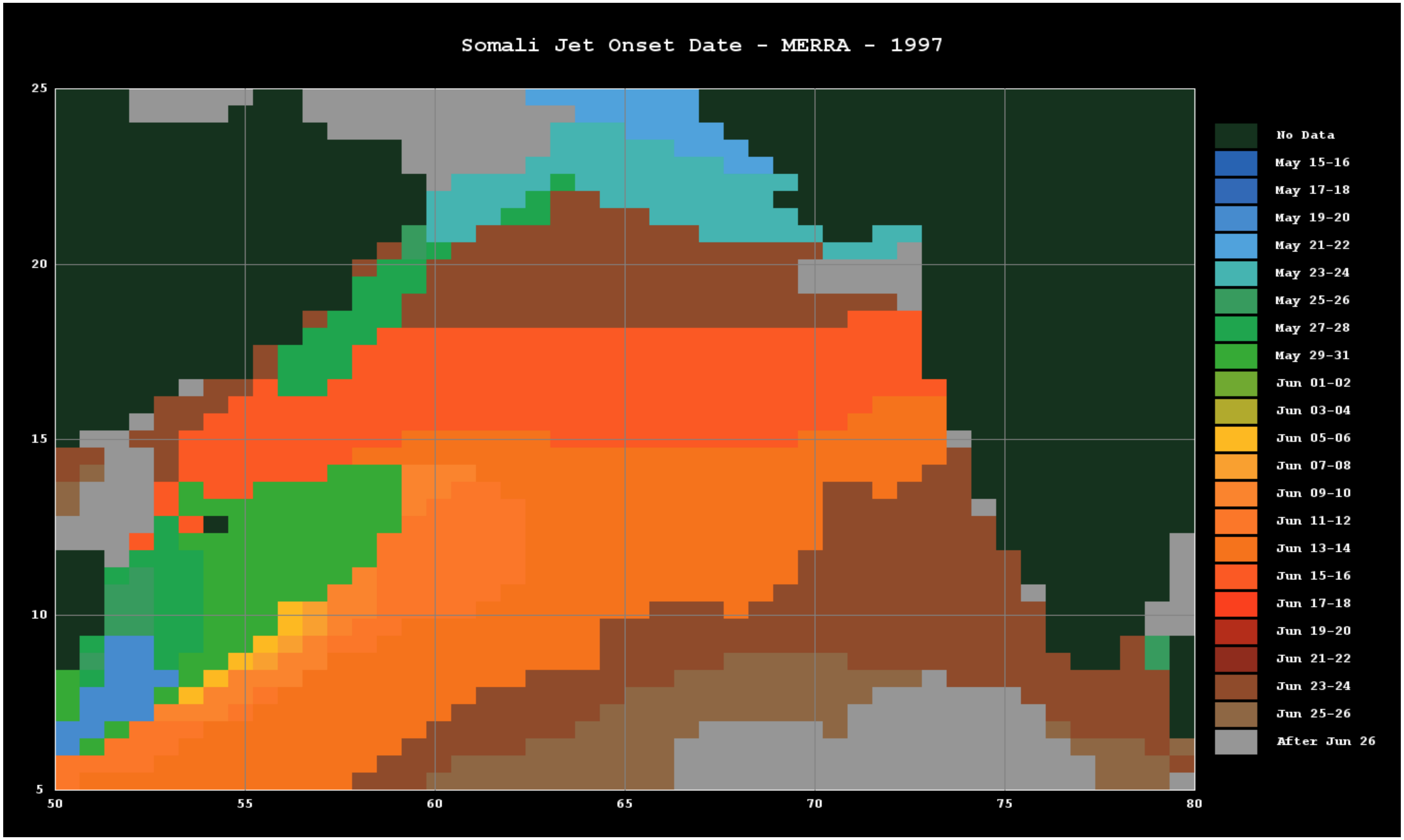


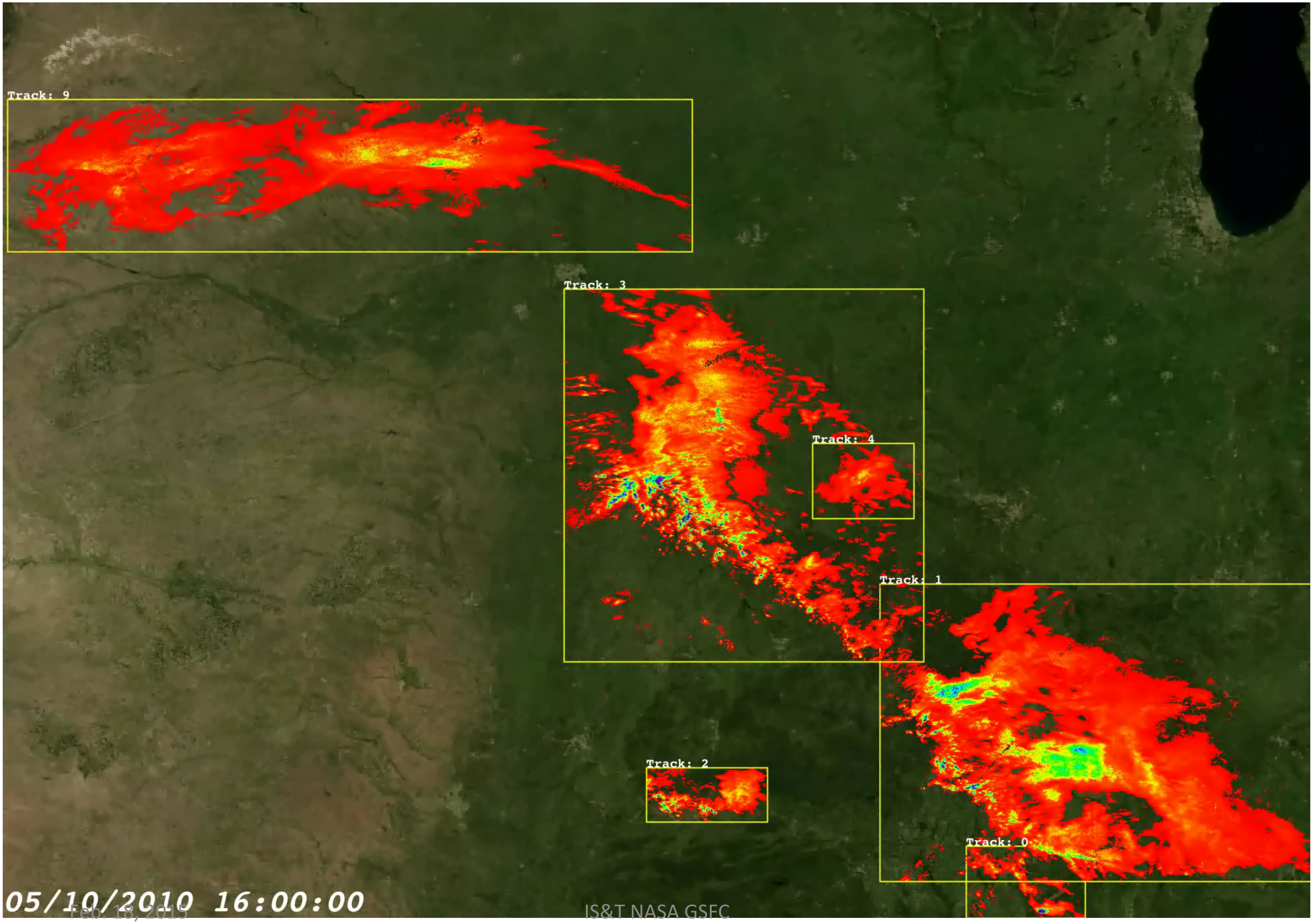




AES generates (lat,lon,time) tables that can be used to access relevant observations (e.g. via ECHO)

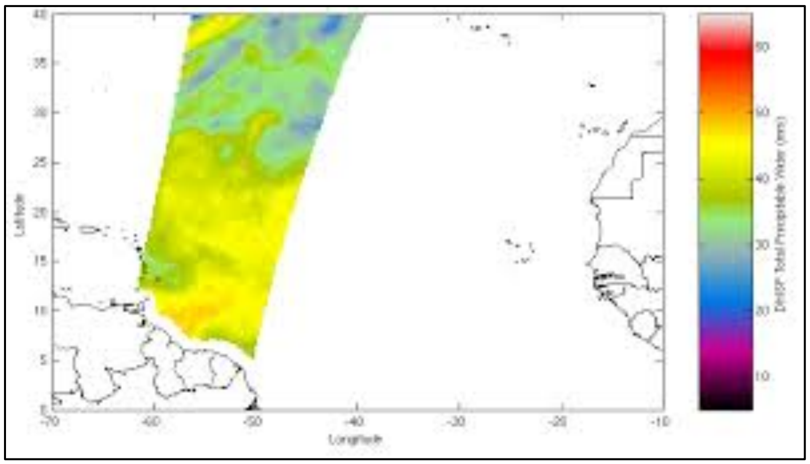
Somali Jet – MERRA Results

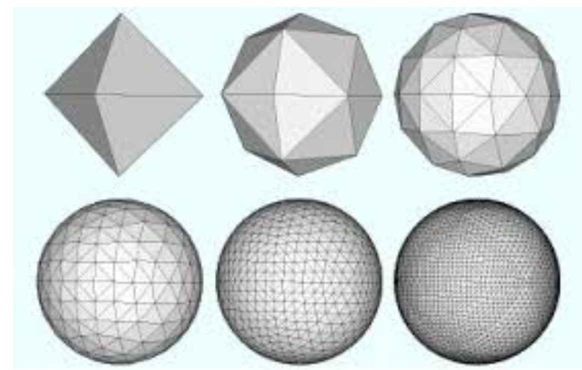


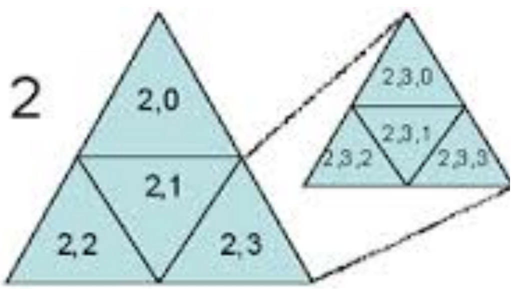

Track: 9
Track: 3
Track: 4
Track: 1
Track: 2
Track: 0
05/10/2010 16:00:00

AIST14: DERECHOs

(1) Improved support for swath and point data:







Hierarchical Triangular Mesh (HTM)

AIST14: DERECHOs

(2) Nonlinear dimensional reduction

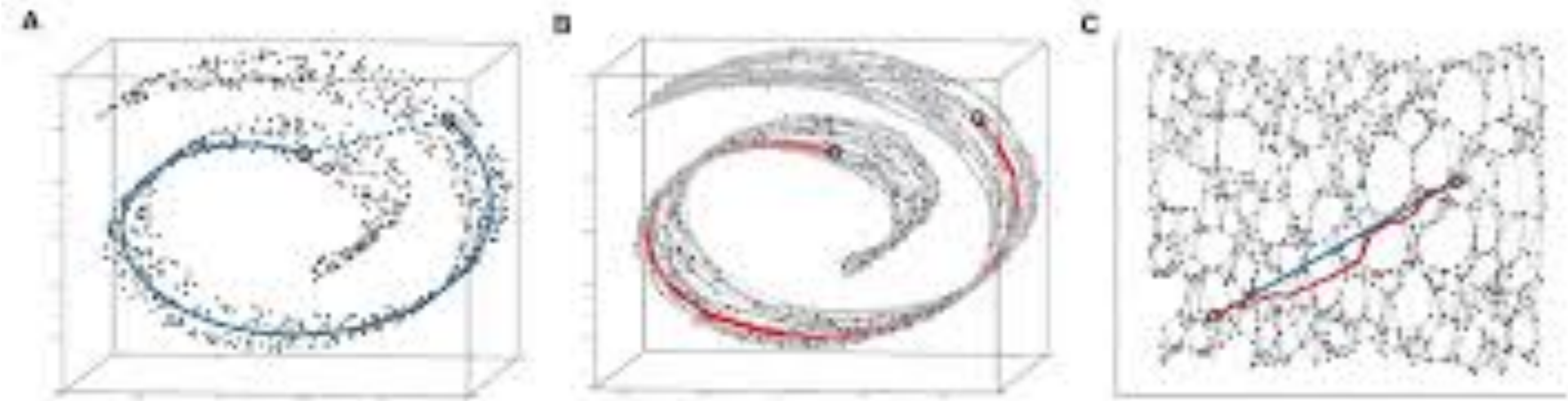


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. **(A)** For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). **(B)** The neighborhood graph G constructed in step one of Isomap (with $K = 7$ and $N = 1000$ data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in G. **(C)** The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).

AIST14: DERECHOs

(3) Moving objects:
Enables reasoning about continuity *between* snapshots in time.

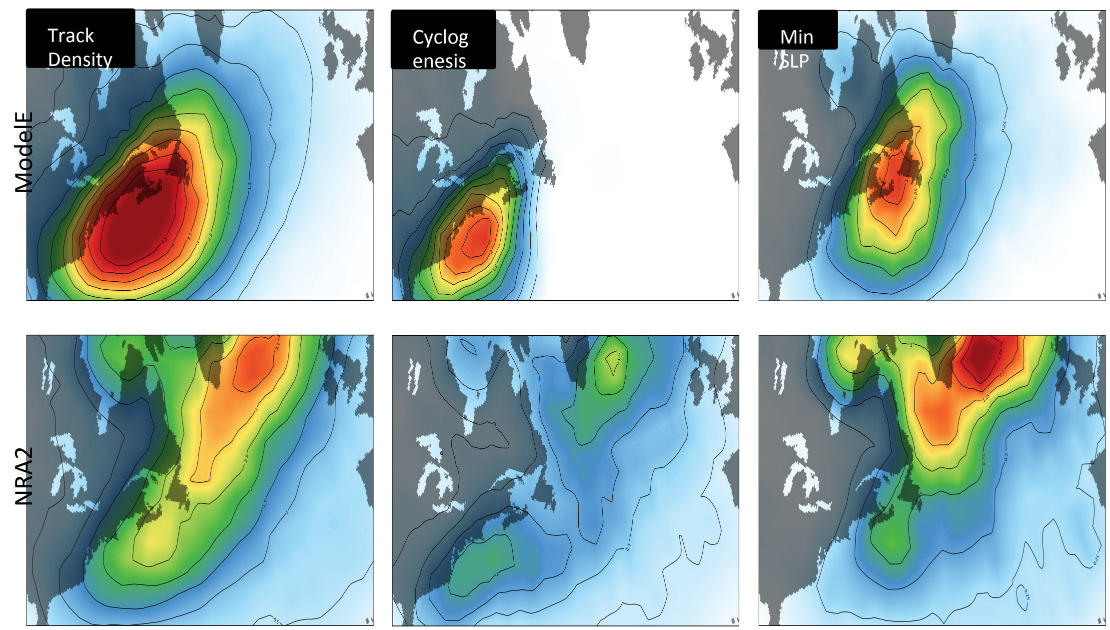
- Where/when did hurricane make landfall?
- Which instruments could see this event?
- ???

Thank you!

Extra material

Associated Projects

- PROBE - Process Based Diagnostics (PBDs)
  - Enable comparison of event statistics between climate models, reanalyses, and observational data.
  - Aid in identification of physical process responsible for divergence.



Associated Projects

- AES in the cloud
  - Explore feasibility of using cloud-based computing and storage for AES deployment
- Advantages:
  - Allows for bursty/intermittent requirements
  - Enable access to external researchers

Somali Jet – SSM/I Results